

Criteria for Modeling Equivalent and Non-Equivalent Phraseological Units in an Electronic Dictionary Database

ISKANDAROVA AYBIBI RAKHIMBAYEVNA
National University of Uzbekistan, Tashkent, Uzbekistan

ABSTRACT

This article presents theoretical considerations on the criteria for creating a linguistic database of translation units from literary works. Identification models for equivalent and non-equivalent variants of phraseological units in Russian-Uzbek novels are proposed for the linguistic database. Identifying phraseological units in an electronic linguistic database, distinguishing them from other language units, and systematizing their structural and semantic properties make it possible to recognize them effectively in automated processing. During identification, encoding the formal structure, grammatical model, lexical meaning, semantic integrity, variants, and functional properties of a phraseological unit with special tags ensures accuracy and consistency in machine translation. For phraseological units occurring in 22,108 aligned segments obtained from two sources, translation pairs were classified as phraseologism-to-phraseologism correspondences or free phrase-to-phraseologism transformations, and corresponding identification models were proposed. Such encoding of phraseological units can improve machine-translation quality by up to 40%.

Keywords: Corpus linguistics, electronic dictionary database, phraseological units, language resources, modeling criteria, linguistic annotation, translation-unit identification, parallel corpus

1. INTRODUCTION

In corpus linguistics, studies of parallel corpora and machine translation frequently describe phraseological units in morphological terms, including the part of speech of the head component, number, possessive and case forms, and possible variants. Such morphological tags facilitate the identification and automatic recognition of phraseological units in machine-translation systems [1; 11].

In contemporary corpus linguistics, modeling phraseological units for linguistic databases is an important stage in the development of language resources. Their linguistic descriptions are formalized, and their lexical, semantic, grammatical, pragmatic, and structural properties are systematized according to explicit criteria. This approach promotes database consistency and improves the efficiency of automatic retrieval, semantic analysis, and natural language processing [2].

The phraseological status of a unit is determined by such features as semantic integrity, component stability, figurative meaning, and reproducibility as a conventional unit in discourse. In machine translation, these properties help distinguish phraseological units from free word combinations [3].

Lexical identification is one of the principal modeling criteria. The head component of an expression is classified as a noun, adjective, verb, or another independent part of speech and entered into the database retrieval system. This procedure provides a structural model of phraseological units. For machine-translation purposes, grammatical patterns such as Noun + Noun, Numeral + Adjective, and Noun + Adverb + Verb are identified so that units can be grouped, retrieved, and automatically classified according to their grammatical structure. At the semantic level, each phraseological unit is recorded not only with its nominative meaning but also with its figurative meaning, emotive connotations, evaluative meaning, and conceptual content. Where relevant, synonyms, antonyms, and membership in a thematic group are also specified, thereby extending the database's semantic-search capabilities.

The stylistic-pragmatic description records, in separate fields, whether an expression occurs in colloquial, literary, or journalistic discourse, together with its degree of expressiveness, emotive connotations, and positive or negative evaluative meaning. These parameters help identify the functional properties of phraseological units. Corpus evidence is another essential component of the modeling process. Each phraseological unit is supported by examples from original sources, and its frequency, distribution across functional styles, and contextual properties are recorded. These corpus data enhance the reliability of the dictionary database and support the development of a structured metadata system. For each phraseological unit, the database records an identification number, lemma, structural model, semantic description, grammatical features, stylistic description, variants, synonyms and antonyms, usage example, source, and corpus frequency. Together, these parameters constitute the conceptual model of the electronic dictionary database [4].

2. RELATED WORK

Raja & Vats emphasize the central role of parallel corpora in the development of machine-translation approaches ranging from statistical machine translation (SMT) to modern neural machine translation (NMT). For low-resource languages in particular, parallel corpora provide essential data for training high-quality translation models and supporting cross-lingual transfer learning, multilingual semantic modeling, and zero-shot translation [5].

Tiedemann notes that the openly available OPUS parallel corpora, covering more than 600 languages and over 40,000 language pairs, have become one of the most important components of multilingual linguistic infrastructure [6]. He further argues that large parallel corpora are key resources not only for training machine-translation systems but also for developing multilingual language models and advancing semantic analysis and other natural language processing tasks.

Under the supervision of N. Abdurakhmonova, the project ‘PARATRANSLATOR: Creating a Contextological Electronic Translation Dictionary Platform Based on a Parallel Corpus’ was

carried out, resulting in an electronic dictionary platform designed to provide context-sensitive translations. The Paratranslator database contains ten languages – Uzbek, English, Russian, Turkish, French, Spanish, Karakalpak, Japanese, Korean, and German—and comprises 960,000 alignment units representing literary, scientific, official, and journalistic discourse [7]. Its Russian-Uzbek literary component includes 8,940 sentences from Part I and 6,589 sentences from Part II of L. Tolstoy's *Anna Karenina*, represented in the Russian original and M. Ismoilov's Uzbek translation, together with 5,000 alignment units based on Ch. Aitmatov's novel *Plakha* ("Плаха") and its Uzbek translation *Qiyomat* by I. G'afurov, 4,584 alignment units from V. Remizov's novellas and short stories translated by M. Omon, and 3,951 alignment units from A. Chekhov's short stories in the original and in translations by H. Akbar. A total of 2,000 phraseological units extracted from these literary works were identified.

3. THEORETICAL AND METHODOLOGICAL FRAMEWORK

A parallel corpus is a large electronic collection of source texts and their translations that is organized, aligned, and annotated with a defined system of linguistic tags for research and computational processing [8; 9]. Large, openly accessible parallel corpora provide the empirical basis required to investigate cross-lingual equivalence and its representation in bilingual dictionaries, thereby supporting evidence-based research in lexicology and lexicography [10].

N. Abdurakhmonova states that identifying concordances in parallel texts is based on the equivalence of words, word combinations, polysemous words, or fixed expressions in the respective languages. In this case, adequate translation does not always justify itself, because the omission of certain words or the addition of a component introduces changes into their lexical and grammatical model. Therefore, for key words in parallel texts, active words and combinations that occur frequently in context and do not undergo a specific change in translation are collected in the database. V. Sibirtseva notes that studies within a parallel

subcorpus, including semantic text referents, tag-cloud generators, and programs such as LF Aligner, are important for studying different translations of the same work, tracing diachronic and stylistic changes in vocabulary, and gaining insight into translator strategies [12]. Creating a linguistic database of phraseological units in an NLP system is a complex, multi-stage process. Sh. Rahmatullayev argues that phraseological units may be structurally equivalent to a word combination or a sentence and that phraseological meaning should be treated not as a syntactic unit (a speech unit), but as a semantic unit (a language unit): “The main part of phraseological units consists of verbal phraseological units. In the lexical composition of phraseological units, in addition to the verbal word-component, words belonging to other parts of speech also occur. Such components are often expressed by nouns, adjectives, and adverbs” [13].

4. RESULTS AND DISCUSSION

Uzbek is an agglutinative Turkic language, whereas Russian is a fusional language belonging to the Slavic branch of the Indo-European family. These typological differences are reflected in translation structures. In Uzbek, dependent constituents are subordinated, directly or indirectly, to the predicate, which typically occurs at the end of the clause, although literary discourse may employ marked word order and inversion. In Russian, the predicate commonly follows the subject, with other constituents occurring in positions determined by syntactic and information-structural factors. These cross-linguistic differences are also reflected in the morphological tagging of phraseological units in Russian-Uzbek literary translations:

1. *Мне плевать – мена бир пул*

RU – *Мне плевать на него с его княжеством.*

UZ – *Uning knyazligi menga bir pul!*

EN – “His principality is worth nothing to me!”

Мне плевать – Pronoun + Verb; *menga bir pul* – Pronoun + Numeral + Noun

In the Uzbek translation of *Anna Karenina*, the Russian expression *мне плевать* is rendered as *menga bir pul*. Morphological tagging reveals a structural difference: the Russian expression follows the Pronoun + Verb pattern, whereas the Uzbek equivalent follows Pronoun + Numeral + Noun. In the linguistic database, these units are classified according to their head component:

```
<source>
<unit lang="ru" type="phraseologism" pos="verb"
model="Pronoun+Verb">мне плевать</unit>
</source>
<target>
<unit lang="uz" type="phraseologism" pos="noun"
model="Pronoun+Numeral+Noun">menga bir pul</unit>
</target>
<equivalence type="functional"/>
```

When phraseological units from closely related language pairs, such as Turkish-Uzbek, Uyghur-Uzbek, or Uzbek-Azerbaijani, are tagged in a parallel corpus, similar morphological patterns frequently emerge because of their structural and lexical proximity. By contrast, translation units from typologically more distant languages often display greater differences in component structure.

2. *Изрядно под хмельком* – *kayfi taroq*

RU – Они уже были изрядно под хмельком и оттого в ожидании оберовского суда над Авдием нелепо улыбались, о чем-то шушукаясь, подталкивая друг друга и перемигиваясь.

UZ – *Ikkovining ham kayfi taroq, Avdiyni Ober qanday tergashini kutib, nimalarnidir pichir-pichir qilishib, bema' nilarcha iljayishar; bir-birlarini turtib qo' yishar; ko' z qisishardi.*

EN – “Both of them were tipsy; while waiting to see how Ober would interrogate Avdiy, they grinned foolishly, whispered among themselves, nudged one another, and winked.”

Изрядно под хмельком – Adverb + Preposition + Noun; *kayfi taroq* – Noun + Noun. The Russian phraseological unit contains three components, whereas its Uzbek equivalent contains two.

The phraseological units are tagged according to their head component as follows:

```
<source>
<unit lang="ru" type="phraseologism" pos="noun"
model="Adverb+Preposition+Noun">Изрядно под
хмельком</unit>
</source>
<target>
<unit lang="uz" type="phraseologism" pos="noun"
model="Noun+Noun">kayfi taroq</unit>
</target>
<equivalence type="functional"/>
```

3. *Белика беда – Ота го‘ри qozixonami?*

RU – *Да не в этом дело. Подумаешь, поедem – не поедem. Белика беда.*

UZ – *Yo‘q, gap bunda emas. Borsak – borarmiz, bormasak – yo‘q. Shunga ham ota go‘ri qozixonami?*

EN – “No, that is not the point. Whether we go or not—what is the big deal?”

Белика беда – Adjective + Noun; *Ота го‘ри qozixonami?* – Noun + Noun + Noun + Particle. The Russian fixed expression *Белика беда* conveys an ironic evaluation of an event or situation as not particularly serious and performs pragmatic functions associated with irony, dismissiveness, and mitigation.

Ота го‘ри qozixonami? is a colloquial Uzbek expression structured as a rhetorical question and likewise indicates that an event or situation should not be treated as excessively serious. Its grammatical pattern is Noun + Noun + Noun + Particle. The Russian and Uzbek units are close in denotative meaning but differ in their sources of imagery. Because the Uzbek expression is grounded in culture-specific imagery, the pair may be

classified as partially equivalent. Both units are tagged according to their head component as follows:

```
<source>
<unit lang="ru" type="phraseologism" pos="noun"
model="Adjective+Noun">Велика беда</unit>
</source>
<target>
<unit lang="uz" type="phraseologism" pos="noun"
model="Noun+Noun+Noun+Particle">Ota go'ri
qozixonami?</unit>
</target>
<equivalence type="partial"/>
```

4. *Ради Христа – Xudo haqqi*

RU – *Левин забежал опять к жене спросить у нее еще раз, простила ли она его за вчерашнюю глупость, и еще затем, чтобы попросить ее, чтобы она, ради Христа, была осторожнее...*

UZ – *Levin xotinining yoniga kechagi bema' niligini kechirganmi-yo' qligini yana bir marta so' ragani, innankeyin, o' ziga ehtiyotroq bo' lib yurishini yana bir qur iltimos qilgani, ayniqsa: «xudo haqqi, xususan bolalardan o' zingni tortibroq yur, tegib ketishadi yo turtib yuborishadi», deb ogohlantirgani chopib chiqib ketgan edi.*

EN – “Levin ran back to his wife once more to ask whether she had forgiven him for yesterday’s foolishness, then to ask her again to be more careful, especially warning her: ‘For God’s sake, keep your distance, particularly from the children; they might bump into you or push you.’”

Ради Христа – Preposition + Noun; *Xudo haqqi* – Noun + Noun
The Russian expression carries emotive and expressive force and is used in requests or pleas to intensify the appeal; structurally, it follows the Preposition + Noun pattern.

Xudo haqqi is a fixed Uzbek phraseological unit with the Noun + Noun pattern. The Russian and Uzbek expressions are close in denotative meaning and are functionally and semantically equivalent, although their imagery derives from

different linguistic and cultural resources. The units are tagged according to their head component as follows:

```
<source>
<unit lang="ru" type="phraseologism" pos="noun"
model="Preposition+Noun">Ради Христа</unit>
</source>
<target>
<unit lang="uz" type="phraseologism" pos="noun"
model="Noun+Noun">Xudo haqqi</unit>
</target>
<equivalence type="functional"/>
```

Russian-Uzbek literary translations contain both equivalent and non-equivalent phraseological correspondences. Parallel-corpus studies have primarily examined equivalent phraseological units and their morphological models [14; 15]. Translation may also involve phraseologization, that is, the rendering of a free source-language combination as a phraseological unit in the target language. For the linguistic database, such source-target pairs are annotated as follows:

1. *Звуки реального мира – olamning butun qiylu qoli*

RU – И тут вновь ворвались в ее сознание звуки реального мира
 UZ – Shunda yana uning qulog'i ochildi, olamning butun qiylu qolini eshita boshladi.
 EN – "Then her ears opened again, and she began to hear all the noise and commotion of the world."

Звуки реального мира is a free word combination in Russian with the grammatical pattern Noun + Adjective + Noun. In the Uzbek translation, it is rendered by the phraseological expression *olamning butun qiylu qoli*, represented here as Noun + Adjective + Noun. The pair is encoded in the linguistic database as follows:

```
<source>
<unit lang="ru" type="free_phrase" pos="noun"
model="Noun+Adjective+Noun">
```

Звуки реального мира
 </unit>
 </source>
 <target>
 <unit lang="uz" type="phraseologism" pos="noun"
 model="Noun+Adjective+Noun">
 Olamning butun qiylu qoli
 </unit>
 </target>
 <equivalence type="functional"/>

2. *To всё равно – ikki dunyo bir qadam*

RU – Я не могу их соединить, а это мне одно нужно. А если этого нет, то все равно.

UZ – Agar qovushtirolmasam – u vaqtda menga ikki dunyo bir qadam.

EN – “If I cannot bring them together, then for me the two worlds are but one step.”

The Russian expression *To всё равно* conveys the meaning “it does not matter” or “it makes no difference” and is rendered in Uzbek as *ikki dunyo bir qadam*. The Russian unit is analyzed as Particle + Pronoun + Predicative, whereas the Uzbek equivalent is a nominal phraseological unit with the pattern Numeral + Noun + Numeral + Noun. Thus, the pair illustrates functional equivalence across different structural types.

<source>
 <unit lang="ru" type="free_phrase" pos="predicative"
 model="Particle+Pronoun+Predicative">
 To всё равно
 </unit>
 </source>
 <target>
 <unit lang="uz" type="phraseologism" pos="noun"
 model="Numeral+Noun+Numeral+Noun">
 Ikki dunyo bir qadam
 </unit>
 </target>
 <equivalence type="functional"/>

3. Он бонвиван большой – *rosa oladigan xilidan-da*

Он бонвиван большой is treated as a noun-headed free combination with the pattern Pronoun + Noun + Adjective and refers to a person inclined toward pleasure and entertainment. In the Uzbek translation, *rosa oladigan xilidan-da* functions as a fixed expression and is analyzed as a nominal phraseological unit with the pattern Adverb + Verb (Participle) + Noun + Particle.

For the linguistic database, the pair is encoded as follows:

```
<source>
<unit lang="ru" type="free_phrase" pos="noun"
model="Pronoun+Noun+Adjective">
Он бонвиван большой
</unit>
</source>
<target>
<unit lang="uz" type="phraseologism" pos="noun"
model="Adverb+Participle+Noun+Particle">
Rosa oladigan xilidan-da
</unit>
</target>
<equivalence type="functional"/>
```

4. Желание иметь детей – *bola tirnog' iga zorlik*

RU – *Его желание иметь детей она объясняла себе тем, что он не дорожил ее красотой.*

UZ – *Vronskiyning bola tirnog' iga zorligini: «U husnimning qadriga yetmaydi» deb tushunardi.*

EN – “She interpreted Vronsky’s longing for a child as: ‘He does not appreciate my beauty.’”

Желание иметь детей is a free Russian word combination with the pattern Noun + Verb (Infinitive) + Noun. Its Uzbek counterpart, *bola tirnog' iga zorlik*, is a nominal phraseological unit characterized by semantic integrity, stability, and figurativeness and is represented by the pattern Noun + Noun + Noun.

```

<source>
<unit lang="ru" type="free_phrase" pos="noun"
model="Noun+Verb (Infinitive)+Noun">
Желание иметь детей
</unit>
</source>
<target>
<unit lang="uz" type="phraseologism" pos="noun"
model="Noun+Noun+Noun">
Bola tirnog'iga zorlik
</unit>
</target>
<equivalence type="functional"/>

```

5. *He имеет смысла – kurakda turmaydigan narsalar*

RU – ... но ведь то, что ты говоришь, или не имеет смысла, или имеет очень дурной смысл.

UZ – ...lekin sen aytgan narsalar yo bema'ni narsalar yoki kurakda turmaydigan narsalar.

EN – "...but what you are saying is either meaningless or utter nonsense."

He имеет смысла is a negative verbal combination indicating that an action, idea, or situation lacks logical or practical justification; its pattern is Particle + Verb + Noun. The Uzbek expression *kurakda turmaydigan narsalar* is a nominal phraseological unit referring to something illogical or unreasonable and follows the pattern Noun + Participle + Noun.

```

<source>
<unit lang="ru" type="free_phrase" pos="verb"
model="Particle+Verb+Noun">
Не имеет смысла
</unit>
</source>
<target>
<unit lang="uz" type="phraseologism" pos="noun"
model="Noun+Participle+Noun">
Kurakda turmaydigan narsalar

```

```

</unit>
</target>
<equivalence type="functional"/>

```

In the electronic dictionary, these pairs can therefore be represented as free_phrase–phraseologism correspondences, marking the contextual transformation of a free source-language combination into a phraseological target-language unit.

5. PRACTICAL APPLICATIONS

Constructing a linguistic database of translation units from a parallel corpus of literary texts is a multi-stage process based on representativeness, high-quality alignment, contextual and multilayer annotation, and the systematic treatment of variable equivalence. The resulting resource supports translation practice while also providing an empirical basis for research in translation studies and digital linguistics.

A parallel-corpus-based linguistic database of literary texts integrates technical, software, and linguistic resources for collecting, storing, processing, retrieving, and analyzing language units [16].

An annotated database of translation units from Russian-Uzbek literary works can contribute to the improvement of automatic translation and to more systematic processing of literary translation data. For translators, it provides evidence for examining translation strategies, identifying semantic differences across languages and cultures, and selecting contextually appropriate equivalents [17].

The linguistic infrastructure of a Russian-Uzbek parallel corpus of literary works requires continuous expansion of its inventory of linguocultural units, including phraseological units, polysemous words, realia, and lacunae. Enriching these resources also creates opportunities for examining historical and semantic developments in both languages.

A parallel-corpus database of translation units is valuable not only as a linguistic resource for machine translation but also as an empirical resource for bilingual lexicography, comparative

translation studies, translation theory, and contrastive linguistics [18].

6. CONCLUSION

Modeling phraseological units for an electronic dictionary database is a complex process based on the integration of linguistic and computer technologies and relies on such criteria as lexical identification, structural description, semantic modeling, variation, stylistic-pragmatic annotation, linking with corpus data, and metadata formation. An electronic dictionary database created on the basis of these criteria provides consistency and practical effectiveness and serves as a reliable linguistic resource for automatic natural-language-processing systems. It contributes to improving the quality of semantic recognition of phraseological units in machine translation. In the linguistic database of translation units from Russian-Uzbek literary works, identifying phraseological units helps determine translation equivalents, monitor their frequency of occurrence in the corpus, compare their contextual realizations, and expand automatic retrieval capabilities.

In conclusion, identifying phraseological units in an electronic linguistic database is one of the principal criteria ensuring the structural integrity, consistency of data, and linguistic value of the database.

REFERENCES

1. Tiedemann, J. 2012. Parallel data, tools and interfaces in OPUS. In *Proc. 8th Int. Conf. Language Resources and Evaluation (LREC)* (pp. 2214-2218).
2. Abdurakhmonova, N. Z. 2018. Linguistic support of the program for translating English texts into Uzbek (on the example of simple sentences). Ph.D. dissertation abstract.
3. Dobrovolsky, D. O. 2020. Корпусный подход к исследованию фразеологии: новые результаты по данным параллельных корпусов. *Вестник Санкт-Петербургского университета. Язык и литература*, 17/3, 398-411.

4. Abdurakhmonova, N. Z., Ismailov, A. S. & Mengliev, D. 2022. Developing NLP tool for linguistic analysis of Turkic languages. In *IEEE International Multi-Conference on Engineering, Computer and Information Sciences – SIBIRCON* (pp. 1790-1793).
5. Raja, R. & Vats, A. 2025. Parallel corpora for machine translation in low-resource Indic languages: A comprehensive review. In *Proceedings of the NAACL 2025 Workshop on Technologies for Machine Translation of Low-Resource Languages*.
6. Tiedemann, J. 2023. Democratizing neural machine translation with OPUS-MT. *Language Resources and Evaluation*. doi: 10.1007/s10579-023-09704-w.
7. PARATRANSLATOR. Contextological electronic translation dictionary platform based on a parallel corpus. Available online: <<https://paratranslator.uz/en/>>.
8. Rahimov, A. 2011. *Fundamentals of Computer Linguistics*. Tashkent.
9. Rakhmonova, A. 2022. Computer Methods for Creating the National Corpus of the Uzbek Language. Ph.D. dissertation in Philology, Tashkent.
10. Dobrovolsky, D. O. Корпус параллельных текстов и сопоставительная лексикология. Available online: <<https://www.researchgate.net/publication>>.
11. Abdurakhmonova, N., Alisher, I. & Sayfulleyeva, R. 2022. MorphUz: morphological analyzer for the Uzbek language. In *7th International Conference on Computer Science and Engineering – UBMK* (pp. 61-66).
12. Sibirtseva, V. 2013. Технология использования параллельного подкорпуса Национального корпуса русского языка и коллекций текстов в обучении иностранным языкам. *Rocznik Instytutu Polsko-Rosyjskiego*, 2/5, 98-110.
13. Rahmatullayev, Sh. 1992. *Phraseological Dictionary of the Uzbek Language*. Tashkent.
14. Musurmankulova, Sh. G. 2024. Linguistic Support of a Corpus of Uzbek and Turkish Phraseological Units. Ph.D. dissertation in Philology, Gulistan.
15. Nurullayeva, O. N. 2026. Foundations for Creating a Parallel Corpus of Uzbek and English Phraseological Units. Ph.D. dissertation in Philology, Tashkent.
16. Dorr, B. J. 1994. Machine translation divergences: A formal description and proposed solution. *Computational Linguistics*, 20/4, 597-635.

17. Nord, C. 2005. *Text Analysis in Translation: Theory, Methodology, and Didactic Application of a Model for Translation-Oriented Text Analysis*. 2nd ed. Amsterdam: Rodopi.
18. Nida, E. A. & Taber, C. R. 1969. *The Theory and Practice of Translation*. Leiden: Brill.

ACKNOWLEDGMENT

The author thanks the scholars who contributed to the development of principles for constructing a linguistic database of translation units from a parallel corpus of literary texts. This research was supported by the National Research Program in Computational Linguistics and Applied Linguistics under Grant IL-7823051524.

ISKANDAROVA AYBIBI RAKHIMBAYEVNA
NATIONAL UNIVERSITY OF UZBEKISTAN,
TASHKENT, UZBEKISTAN.